

Parallel Implementation of Efficient Search Schemes for the Inference of Cancer Progression Models

Daniele Ramazzotti^{*†}, Marco S. Nobile^{*‡}, Paolo Cazzaniga^{‡§}, Giancarlo Mauri^{*†} and Marco Antoniotti^{*}

^{*}Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy

Email: daniele.ramazzotti/nobile/mauri/antoniotti@disco.unimib.it

[†]Department of Pathology, Stanford University, California 94305, USA

[‡]SYSBIO.IT Centre of Systems Biology, Milano, Italy

[§]Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy

Email: paolo.cazzaniga@unibg.it

Abstract—The emergence and development of cancer is a consequence of the accumulation over time of genomic mutations involving a specific set of genes, which provides the cancer clones with a functional selective advantage. In this work, we model the order of accumulation of such mutations during the progression, which eventually leads to the disease, by means of probabilistic graphic models, i.e., Bayesian Networks (BNs). We investigate how to perform the task of learning the structure of such BNs, according to experimental evidence, adopting a global optimization meta-heuristics. In particular, in this work we rely on Genetic Algorithms, and to strongly reduce the execution time of the inference—which can also involve multiple repetitions to collect statistically significant assessments of the data—we distribute the calculations using both multi-threading and a multi-node architecture. The results show that our approach is characterized by good accuracy and specificity; we also demonstrate its feasibility, thanks to a $84\times$ reduction of the overall execution time with respect to a traditional sequential implementation.

I. INTRODUCTION

Cancer development is driven by the subsequent accumulation of genomic mutations over a set of *driver* genes, which confer a functional selective advantage to the cancer clones, leading to the emergence and further development of the disease. Indeed, during clonal expansions, tumor cells compete for space and resources and only the fittest clones are capable of outgrowing the competing cells [1], [2]. Here, we aim at modeling such systems in terms of dynamic processes of monotonic accumulations of driver alterations over time.

Bayesian Networks (BNs) can be exploited to describe the dynamics of biological phenomena characterized by the monotonic accumulation of events, e.g., gene mutations in the context of cancer progression [3]. Following this methodology, a temporal ordering among events is implied, so that the occurrence of an early event positively correlates with the subsequent occurrence of its successors. Such an approach has recently been proved to be effective in [4], where progression models are inferred on a cohort of patients derived from two subtypes of colorectal tumors.

In this work, we first focus on the problem of inferring a BN—which is a well-known NP-hard problem [5]—along with the characterization of its complexity and pitfalls. Then, we describe possible heuristics to perform the inference of the

network and present an efficient parallel implementation of this procedure based on Genetic Algorithms (GAs). Finally, we present the results obtained by the application of this approach to synthetic data generated by realistic statistical models, pointing out the satisfactory performance in terms of structural distance from the generative synthetic ground truth and the improvement in execution time achieved thanks to the proposed parallel implementation.

The paper is structured as follows. In Section II we describe the problem of the BN inference for the specific problem of cancer progression. We then describe the strategy adopted to tackle this problem (GAs), as well as the parallel architecture used to accelerate the inference process. In Section III we present the results obtained from the inference process and the computational speed-up achieved. We conclude with some remarks and future developments of this work.

II. METHODS

A. Bayesian Networks and Cancer Progression

Bayesian Networks are probabilistic graphical models, encoded as Directed Acyclic Graphs (DAGs), which describe the conditional dependence relations among random variables. Formally, a BN is defined as a DAG $G = (V, E)$, where V is the set of nodes representing any considered random variable, and E is the set of arcs describing the conditional dependencies among nodes [6].

Recently, several statistical methods have been proposed to exploit such Bayesian graphical models to the aim of describing the evolution and development of cancer progression in terms of accumulation of genomic mutations over time (see an example in Figure 1). In such a case, V represents the set of genomic mutations, while E represents the preferential ordering of accumulations among such mutations, depicted as relations of selective advantage in the graph [3], [7], [8]. A progression model can also be encoded by using an adjacency matrix, which will be exploited here during the inference process, where the value 1 denotes that the genomic mutation of a given row favors the genomic mutation of the corresponding column. Table I reports the binary adjacency matrix of the BN shown in Figure 1.

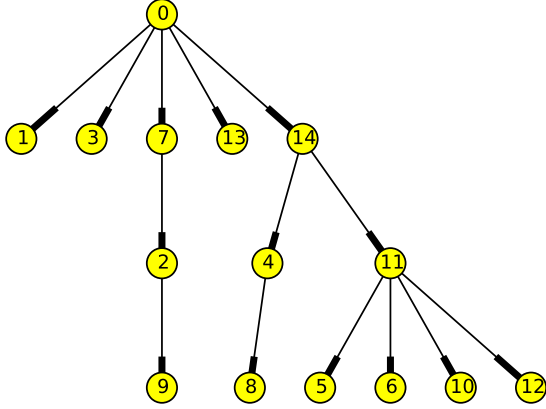


Fig. 1. Example of a cancer progression model. The nodes in the graph represent gene mutations, while the arcs describe the conditional dependency between the mutations.

The input of the statistical methods used for the inference is an additional binary matrix $\mathbf{O} \in \{0, 1\}^{M \times K}$ that contains the experimental observations. This matrix is composed by M rows, one for each observation (e.g., genomic data of a patient affected by cancer), and by $K = |V|$ columns, which represent the genes whose progression has to be inferred. A value $o_{m,k} \in \mathbf{O}$ (with $m = 1, \dots, M$ and $k = 1, \dots, K$) is set to 1 when gene k is mutated in the biological sample related to the m -th patient; on the contrary, $o_{m,k} = 0$ to denote the absence of such mutation.

TABLE I
MATRIX REPRESENTATION OF A BN.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0

Furthermore, state-of-the-art techniques that make use of BNs to model cancer progression, apply further constraints to the obtained network, leveraging the theory of probabilistic causation introduced by Patrick Suppes [9]. Specifically, Suppes defined the notion of *prima facie causation*, where a relation between a cause u and its effect v is verified when two conditions are observed to be true:

- 1) *temporal priority* (TP): each cause must precede its effects;
- 2) *probability raising* (PR): the presence of the causing event increases the probability of observing its subsequent effects.

Given these conditions, we can formulate the following definition [3], [9]:

Definition 1 (Probabilistic causation [9]): For any couple of events u and v , occurring respectively at times t_u and t_v , under the mild assumptions that $0 < P(u), P(v) < 1$, the event u is called a *prima facie cause* of v if it occurs *before* u and it *raises the probability* of u :

$$\begin{cases} (TP) & t_u < t_v \\ (PR) & P(v|u) > P(v|\neg u). \end{cases} \quad (1)$$

Hence, the results of the statistical methods are constrained Bayesian networks—named Suppes-Bayes Causal Networks (SBCNs) [3]—which account for the selective advantage relations among genomic events by combining probabilistic constraints with maximum likelihood estimation [9]. Therefore, the problem of determining the sequence of mutations leading to cancer can be re-formulated as the problem of learning the structure of a BN.

There exist two strategies to tackle this structure inference problem: (i) the constraint-based approaches, mainly due to the works of Judea Pearl, which consist in discovering the conditional independence relations within the input data to learn the BN [6]; (ii) the score-based approaches, in which the inference problem is re-stated as an optimization problem where all possible DAGs are considered valid solutions and they are evaluated using a likelihood-based score function [6].

In this paper, we rely on the latter strategy. This approach assumes the data to be independent and identically distributed and, because of this, the likelihood of the data is the product of the likelihood of each datum, which, in turn, is defined by the factorized joint probability function described as follows:

$$P(x_1, \dots, x_k) = \prod_{X_i \in V} P(X_i = x_i | Pa(X_i) = x_{Pa(i)}),$$

where x_1, \dots, x_k are the nodes in the network and $Pa(\cdot)$ indicates the parent set (i.e., all the nodes with an arch pointing to it) of a given node.

For numerical reasons, log-likelihood (LL) is usually used instead of the likelihood itself, and thus the likelihood product becomes the log likelihood sum. To avoid overfitting, we use the so called Bayesian Information Criterion (BIC) [10] “regularized” likelihood score function (which is calculated using the package `bnlearn` of the R programming tool [11]). The score function is defined as follows:

$$score_{BIC}(D, G) = LL(D|G) - \frac{\log M}{2} \dim(G), \quad (2)$$

where G denotes the considered DAG, D denotes the input data, M denotes the number of samples, and $\dim(G)$ denotes the number of parameters of the DAG G . Here, the parameters of G refer to the set of arcs present in G along with the encoded conditional probabilities. It is worth noting that the regularization term $-\dim(G)$ is exploited to favor nodes characterized by fewer parents, so that sparse graphs are promoted during the inference process. This approach is adopted with the aim of providing models with the most confident set of arcs, even admitting the possibility of missing true relations.

In fact, intuitively, we add to the inferred model only those arcs that strongly contribute to the calculation of the adopted likelihood score, which, in terms of likelihood, are also the most confident ones.

As the size of the sample employed in the inference process increases, both the weight of the regularization term and the weight of the likelihood function increase. However, the increment of the latter is more relevant, so that with more input data the likelihood has a more pronounced contribution to the overall score function. Statistically speaking, we can state that the BIC is a consistent score [6], which means that having sufficiently large sample sizes, the network with the maximum BIC score is I -equivalent to the true generative structure [6]. Specifically, two structures are said to be I -equivalent (where I stands for independence), if they encode the exact same set of conditional independence relations among the considered variables. We remark that BNs with different structures may encode the same set of relations of dependency among variables. For this reason, BIC is said to converge to a solution I -equivalent to the ones generating the data, even if there is no guarantee to converge to the exact structure [6].

In general, independently from the strategy used in the inference process, the (huge) size of search space of valid solutions makes this problem very hard to tackle, especially from the computational time point of view.

In particular, as stated above, the problem of BNs inference is NP-hard [5]; therefore all state-of-the-art techniques rely on heuristics [6], and they are mainly based on stochastic population-based global optimization algorithms. Specifically, methods based on Genetic Algorithms [12] and Ant Colony Optimization [13]–[15] have been proposed in the literature. In this work we exploit the first methodology, described in the next section, and we investigate how the chosen heuristics impacts the performance.

B. Genetic Algorithms and Model Inference

Genetic Algorithms (GAs) were introduced by J. H. Holland in 1975 [16] as a global search methodology inspired by the mechanisms of natural selection. In GA, a population \mathcal{P} of candidate solutions iteratively evolves towards the global optimum of a user-specified fitness function.

GAs are characterized by a well-known convergence theorem named *schema theorem*. This theorem ensures that the presence of a schema (i.e., a template of solutions) in the population, having a good impact on the fitness value (i.e., the quality of a candidate solution), increases exponentially generation after generation. GAs were shown to be effective for the problem of the BN learning, both in the case of available and not available *a priori* knowledge about nodes' ordering [12], which allows a relevant reduction of the search space.

GAs are based on a population \mathcal{P} composed of Q randomly created individuals that are generally represented as fixed-length strings over a finite alphabet, encoding solutions of the problem under investigation. In this work, each individual

represents the linearized adjacency matrix of a candidate BN (e.g., the matrix in Table I), by means of a string of binary values whose length is $K \times K$.

The individuals of the population undergo an iterative process whereby three genetic operators (selection, crossover, mutation) are applied, according to a given fitness function, to simulate the evolution process which results in a new population of possibly improved solutions. The fitness function used in this work is the score formalized in Equation 2. During the selection process, individuals from \mathcal{P} are chosen and inserted into a new temporary population \mathcal{P}' using some fitness-dependent sampling procedure [17]. In this work we assume a ranking selection: individuals are ranked according to their fitness values and the probability of selecting an individual is proportional to its *position* in the ranking.

The crossover operator is used to combine the structure of two promising parents into new and improved offspring, which are collected into a third population \mathcal{P}'' . We assume a single point crossover, in which the two strings encoded by the two parents are “cut” in the same random position and one of the resulting substrings is exchanged. The crossover is generally applied with a probability p_c ; specifically, in our implementation, the crossover is performed on the selected population with $p_c = 1$, although it does not have an actual effect on the offspring when the cut position is equal to 0 or equal to K .

Finally, the mutation operator is used to perturb the solutions encoded in the individuals of the population \mathcal{P}'' , thus allowing a further exploration of the search space. Mutation alters a symbol of the individual, which is substituted with a random symbol from the alphabet with a fixed probability p_m . In our tests mutation is applied by flipping a single bit of the individual with probability $p_m = 0.01$, as suggested in [12].

After the application of the genetic operators, GAs can proceed by following two alternative strategies: (i) all individuals in \mathcal{P}'' replace those in \mathcal{P} (simple method); (ii) the best Q individuals in $\mathcal{P} \cup \mathcal{P}''$ replace those in \mathcal{P} (elitist method, which is exploited in this work). Once the population is replaced, the process iterates until a halting criterion is met, e.g., after a fixed number of generations (100 in this work).

It is worth noting that the one-point crossover and the mutation are not closed operators in the case of unordered nodes, since the resulting offspring might not encode valid DAGs. To the aim of ensuring a consistent population of individuals throughout the generations, the two operators are followed by a correction procedure, in which we analyze the candidate BN to verify the potential presence of cycles. Our correction procedure works as follows: random arcs are removed from the solution until no more cycles are detected. Cycles in the networks are detected using the `networkx` library [18]; we exploit in particular the `simple_cycles()` method, which returns a list of elementary circuits in the network identified using the Johnson's algorithm [19]. As long as the list is non-empty, we sample and remove arbitrary arcs. The obtained individual (i.e., a DAG) is finally added to the provisional population.

The number of fitness evaluations required by the overall evolutionary methodology—which is proportional to the number of generations and to the size of the population—can be very high, thus resulting computationally challenging. In order to mitigate this problem, our methodology was designed to exploit a high-performance architecture, described in the next section.

C. Distributed Computing on GALILEO

GALILEO is a Tier-1 supercomputer maintained by the Italian Consortium CINECA, devoted to scientific investigation. This supercomputer is composed of 516 computing nodes; each computing node is equipped with two 8-core Haswell processors (clock frequency 2.4 GHz), for a total of 16 cores per computing node and an overall count of 8256 cores. It is worth noting that hyperthreading is disabled on this computing nodes, reducing the overall level of parallelism to the physical 8 cores. Computing nodes are also equipped with 128 GB of RAM (8 GB reserved for each core). GALILEO also contains 768 Intel Xeon Phi 7120p co-processors and 20 GPUs Nvidia K80, although none of them was exploited in this work.

Our learning algorithm was implemented in a multi-threaded fashion, to fully leverage the cores of the computing nodes and distribute the fitness evaluations required by the GA. As a further acceleration, we exploited the MPI library [20] to distribute over several computing nodes the execution of multiple parallel optimizations, whose results are used to calculate the statistical data about the learning process. As a last note, we point out that GALILEO's job scheduling system limits to 100 the number of simultaneous computing nodes that can be requested for a single job. Hence, we subdivided all tests into separate jobs composed of 100 simultaneous optimizations.

III. RESULTS

We tested the performance of our method both in terms of error rates between the inferred structure and the true one (i.e., the exact structure), and in terms of speed-up of the running time achieved with respect to a strictly sequential execution. To this aim, we first generated a set of random structures, used as ground truth, representing BNs that are used as generative models for *in silico* observations of genomic profiles. Given these BNs, we sampled a set of datasets used as the starting point to perform the learning task. In order to mimic cancer progression and, specifically, its cumulative dynamics, we also constrained the conditional probabilities of the randomly generated BNs so that they only model positive dependencies among nodes. Stated in other words, when generating the random structures to be learned, the random BNs model only situations where the presence of a parent node (positively) correlates (i.e., increases) the expected probability of later observing its child. So doing, we describe probabilistic relations of selective advantage among cancer clones, where the occurrence of an early mutation increases our expectation of observing, later on, its subsequent mutation during cancer progression [8].

A. Inference Results

We first tested the performance of our method in terms of structural distance of the inferred solutions from the generative model (i.e., the exact structure). The performance is evaluated using the classic measures of accuracy, sensitivity and specificity defined as follows:

- $accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$;
- $sensitivity = \frac{TP}{(TP+FN)}$;
- $specificity = \frac{TN}{(FP+TN)}$;

where TP and FP denote true and false positives, respectively, while TN and FN are the true and false negatives, respectively. We define as positive any arc that is present in the generative model, while negative any arc that is not present in the generative model. To be more precise, TP are the arcs present in the generative model and correctly inferred by the method, while FP are the arcs not inferred but present in the ground truth model. On the contrary, TN are the arcs not present in the generative model and (correctly) not inferred by the method, while FN are the arcs not inferred but present in the ground truth model. The measures of accuracy, sensitivity and specificity assume real values in $[0,1]$; a result ≈ 1 indicates a good performance of our inference method.

To assess the performance of our method we generated *in silico* 200 random trees composed of 15 nodes (in this test we assign at most one predecessor per node). Note that the size of random trees employed in our tests is analogous to the size of real models of cancer progression (see, for instance, [21]). Moreover, to further complicate the problem, the 200 random trees were subdivided in sets of 40 synthetic causal trees characterized by 5 increasing levels of noise (0%, 5%, 10%, 15% and 20%). Specifically, we applied a number of random perturbations according to the noise level to the observation matrices (i.e., bit flipping) to model any potential source of error that can occur during the experimental data collection.

Finally, for each tree, we generated a dataset of $M = 100$ observations, encoded as binary matrices. We then executed the test of performances both for the reconstruction of constrained and unconstrained BNs (as described in [8]). We want to point out that the methodology described in this paper is not limited to the inference of trees as it can, in principle, reconstruct any DAG [7], [12].

Tables II and III reports the results obtained for the inference of unconstrained BNs and SBCNs, respectively. We observe high values of accuracy and specificity in both cases, which remain above 0.9 in all conditions, demonstrating how our methodology is also robust to different levels of noise in the observed data. In the case of SBCNs, however, we achieved better results, reaching a value of 0.97 for the accuracy in the case of noise-free observations. SBCNs are also characterized by a value of 0.98 for the specificity (i.e., capability of avoiding false positives) also assuming a relevant level of noise. We observe that the values of sensitivity obtained in these tests are consistent with the expected performance of the adopted regularizer (i.e., BIC), which is specifically

adopted here to converge to the most confident relations among genes. As a matter of fact, BIC is designed to produce sparse graphs, involving only the most likely arcs [3]. For this reason, the sensitivity values (a measure of false negatives) are typically lower than the ones of specificity (a measure of false positives), as BIC (at least with small sample sizes) tends to provide sparse networks that well describe the data, rather than predicting more arcs. If the goal was to perform prediction of possible unknown relations among genes while admitting false positives, other regularization such as the Akaike information criterion (AIC) [22], may better fulfill this task.

TABLE II

ACCURACY, SENSITIVITY AND SPECIFICITY OF THE GA APPLIED TO THE PROBLEM OF CANCER PROGRESSION INFERENCE TACKLED BY MEANS OF BAYESIAN NETWORKS.

	Noise				
	0%	5%	10%	15%	20%
Accuracy	0.92	0.92	0.93	0.92	0.92
Sensitivity	0.49	0.49	0.55	0.51	0.50
Specificity	0.96	0.96	0.96	0.96	0.96

TABLE III

ACCURACY, SENSITIVITY AND SPECIFICITY OF THE GA APPLIED TO THE PROBLEM OF CANCER PROGRESSION INFERENCE TACKLED BY MEANS OF SUPPES-BAYES CAUSAL NETWORKS.

	Noise				
	0%	5%	10%	15%	20%
Accuracy	0.97	0.97	0.96	0.95	0.92
Sensitivity	0.84	0.83	0.76	0.71	0.58
Specificity	0.98	0.98	0.98	0.98	0.95

B. Computational Performances

GAs belong to the class of iterative and population-based optimization meta-heuristics. Thus, during each generation, the fitness evaluations must be calculated for each individual. Since all individuals are mutually independent, the process of fitness evaluations can be parallelized.

In this work, the fitness evaluations—which are based on BIC—are performed using the package `bnlearn` of the R programming tool [11]. Multiple threads are created, each one running an instance of `bnlearn`, to independently calculate the fitness values. A join mechanism allows to synchronize the termination of all threads, collect the results and update the fitness values for all individuals at once. By using this multi-threaded strategy we obtained a speed-up—with respect to a strictly sequential execution on the same computing node—approximately equal to $8\times$ with a populations \mathcal{P} of the GA with a number of individuals $Q > 8$ (see Figure 2). For instance, in the case of the test with $Q = 128$ individuals, the running time of the optimization was reduced from 1 hours and 22 minutes down to 647 seconds, corresponding to a $7.6\times$ speed-up.

The second level of parallelism was introduced by means of MPI, which was used to distribute the GA instances over multiple computing nodes. The results in Figure 3 summarize

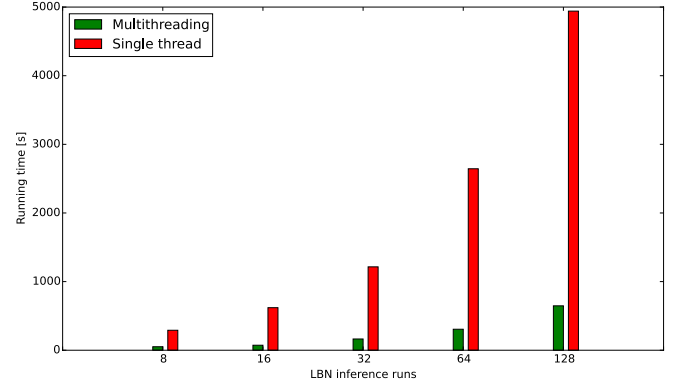


Fig. 2. Comparison of the running time (in seconds) of the learning algorithm using distributed fitness evaluations (yellow bars) with respect to a classic single threaded execution (red bars).

the speed-up obtained using this approach, highlighting that for a few optimizations the overhead due to MPI reduces the potential acceleration. However, when the number of parallel optimizations is greater than 10, the distributed architecture strongly reduces the execution time of the inference processes. Not surprisingly, the maximum acceleration is achieved in the case of 100 simultaneous GA executions, completed in 24 minutes, where the same job executed on a single node takes 4 hours and a half, corresponding to a $11\times$ speed-up. Clearly, the restriction on the number of computing nodes that can be concurrently employed on GALILEO supercomputer limits the performance gain that can be achieved. Indeed, in the case of 200 optimizations we still achieved an overall speed-up equal to $11\times$ (see Figure 3).

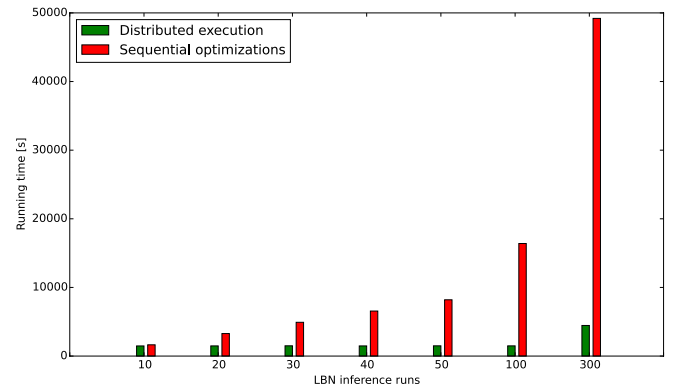


Fig. 3. Comparison of the running time (in seconds) of the learning algorithm distributed on the GALILEO supercomputer (yellow bars) with respect to a simple multi-threaded execution (red bars).

In conclusion, the relevant reduction of the running time provided by supercomputers can be exploited to assess the

confidence in the estimation of an inferred network. As a matter of fact, as discussed in details in [4], the statistic support of a learned network can be calculated by performing multiple optimizations and by repeatedly sampling the input data (namely, bootstrap or cross-validation approaches). In this case, it is suggested to perform at least 100 repetitions of the optimization process, which would benefit from our distributed approach.

IV. CONCLUSION

In this work we presented a methodology for the efficient inference of models of cancer progression from genomic data, and we assessed its performances in terms of accuracy, sensitivity and specificity, showing the good results related to the inferred BNs and the robustness of the optimization process in the case of noisy experimental data.

The methodology exploited in this paper is based on Genetic Algorithms [12], and it is accelerated by means of the combination of multi-threading and distributed computation. Thanks to our approach, the overall computation time was reduced of almost two orders of magnitude using a parallel architecture (CINECA's GALILEO): on the one hand, the multi-threaded execution of the fitness functions allowed a $7.6\times$ speed-up on each computing node; on the other hand, the parallel execution of multiple optimizations distributed over independent nodes allowed a further $11\times$ speed-up, for an overall reduction of the execution time of approximately $84\times$.

It is worth noting that GALILEO is equipped with additional parallel co-processors, namely GPUs and MICs. Both architectures are characterized by a theoretical peak power of about one tera-flop. By further distributing the calculations on these co-processors, we could strongly reduce the execution time. In particular, GPUs typically contain thousands computing cores that can be used to calculate in a parallel fashion the fitness functions required by the GA, allowing to deal with larger populations of individuals whose optimization would require a reduced running time with respect to the parallel strategy employed in this paper (we refer the interested reader to [23] for a review of GPU-powered methods for computational biology). Thus, as future development of this work, we plan to port the overall methodology to the CUDA architecture.

In our optimization method, both mutation and crossover are non closed operators in the case of non-ordered nodes. This means that they might introduce cycles in the BN, which are identified using Johnson's algorithm and corrected by means of random arcs removal. The complexity of the cycle finding algorithm is $\mathcal{O}((v+e)(c+1))$, where $v = |V|$, $e = |E|$ and c is the number of elementary circuits in the graph. However, all these computations are not necessary, because we just need to know that *at least* one cycle exists in the BN. We will therefore modify the cycle finding algorithm to improve the efficiency of the correction step exploited in this work.

ACKNOWLEDGMENT

The authors would like to thank A. Tangherloni for his suggestions about MPI development. We acknowledge the

CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

REFERENCES

- [1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.
- [2] —, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.
- [3] D. Ramazzotti, A. Graudenzi, and M. Antoniotti, "Modeling cumulative biological phenomena with Suppes-Bayes causal networks," *arXiv preprint arXiv:1602.07857*, 2016.
- [4] G. Caravagna, A. Graudenzi, D. Ramazzotti, R. Sanz-Pamplona, L. De Sano, G. Mauri, V. Moreno, M. Antoniotti, and B. Mishra, "Algorithmic methods to infer the evolutionary trajectories in cancer progression," *PNAS*, 2016.
- [5] D. M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *The Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.
- [6] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.
- [7] L. O. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti, and B. Mishra, "Inferring tree causal models of cancer progression with probability raising," *PLoS ONE*, vol. 9, no. 10, p. e108358, 2014.
- [8] D. Ramazzotti, G. Caravagna, L. O. Loohuis, A. Graudenzi, I. Korsunsky, G. Mauri, M. Antoniotti, and B. Mishra, "CAPRI: efficient inference of cancer progression models from cross-sectional data," *Bioinformatics*, vol. 31, no. 18, pp. 3016–3026, 2015.
- [9] P. Suppes, *A probabilistic theory of causality*. North Holland, Amsterdam, 1970.
- [10] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [11] M. Scutari, "Learning bayesian networks with the bnlearn R package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.
- [12] P. Larrañaga, M. Poza, Y. Yurramendi, R. H. Murga, and C. M. Kuijpers, "Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 9, pp. 912–926, 1996.
- [13] L. M. De Campos, J. M. Fernandez-Luna, J. A. Gámez, and J. M. Puerta, "Ant colony optimization for learning Bayesian networks," *International Journal of Approximate Reasoning*, vol. 31, no. 3, pp. 291–311, 2002.
- [14] L. M. De Campos, J. Puerta, and J. M. Puerta Castellón, "Learning Bayesian networks by ant colony optimisation: searching in two different spaces," *Mathware & Soft Computing*, vol. 9, no. 3, pp. 251–268, 2008.
- [15] J. Jun-Zhong, H.-X. Zhang, H. Ren-Bing, and L. Chun-Nian, "A Bayesian network learning algorithm based on independence test and ant colony optimization," *Acta Automatica Sinica*, vol. 35, no. 3, pp. 281–288, 2009.
- [16] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, Michigan, USA: The University of Michigan Press, 1975.
- [17] T. Bäck, "Selective pressure in evolutionary algorithms: A characterization of selection mechanisms," in *Proceedings of the First IEEE Conference on Evolutionary Computation*, vol. 1. IEEE, 1994, pp. 57–62.
- [18] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, 2008, pp. 11–15.
- [19] D. B. Johnson, "Finding all the elementary circuits of a directed graph," *SIAM Journal on Computing*, vol. 4, no. 1, pp. 77–84, 1975.
- [20] L. Dalcín, R. Paz, and M. Storti, "MPI for Python," *Journal of Parallel and Distributed Computing*, vol. 65, no. 9, pp. 1108–1115, 2005.
- [21] M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein, and N. Beerenwinkel, "The temporal order of genetic and pathway alterations in tumorigenesis," *PLoS ONE*, vol. 6, no. 11, pp. 1–9, 11 2011. [Online]. Available: <http://dx.doi.org/10.1371/journal.pone.0027136>
- [22] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 199–213.
- [23] M. S. Nobile, P. Cazzaniga, A. Tangherloni, and D. Besozzi, "Graphics Processing Units in Bioinformatics, Computational Biology and Systems Biology," *Briefings in Bioinformatics*, 2016.